Concept-aware Data Construction
Improves In-context Learning of Language Models

Michal Štefánik[♣][*] and **Marek Kadlčík**[♣] and **Petr Sojka**[♣]

[♣]Faculty of Informatics
Masaryk University, Czech Republic

Abstract

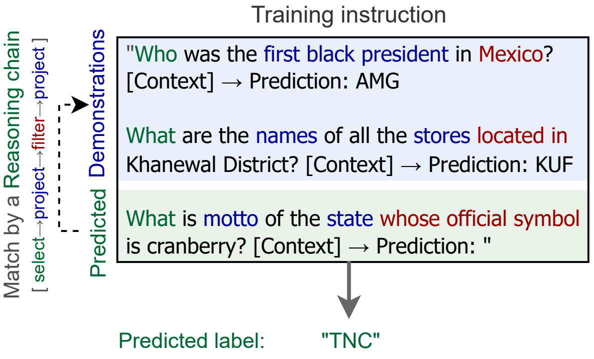
Many recent language models (LMs) are capable of *in-context learning* (ICL), manifested in the LMs' ability to perform a new task solely from a natural-language instruction. Previous work curating in-context learners assumes that ICL emerges from a vast over-parametrization or the scale of multi-task training. However, recent theoretical work attributes the ICL ability to *concept-dependent* training data and creates functional in-context learners even in small-scale, synthetic settings.

In this work, we practically explore this newly identified axis of ICL quality. We propose **Concept-aware Training (CoAT)**, a framework for constructing training scenarios that make it beneficial for the LM to learn to utilize the analogical reasoning concepts from demonstrations. We find that by using CoAT, pre-trained transformers *can* learn to better utilise new latent concepts from demonstrations and that such ability makes ICL more robust to the functional deficiencies of the previous models.

Finally, we show that concept-aware in-context learners are much more effective in in-context learning a majority of unseen tasks compared to traditional instruction tuning, and fare comparably also to previous in-context learners trained in large-scale multitask learning requiring magnitudes of more training data.

1 Introduction

The in-context learning (ICL), as initially uncovered by Brown et al. (2020), is a setting requiring language models (LMs) to infer and apply correct functional relationships from the pairs of inputs and outputs (i.e. *demonstrations*) presented in user-provided input prompt (Li et al., 2023b). Given that a small set of demonstrations can be obtained for any machine learning task, in-context learning presents a much more versatile and practical alternative to training task-specific models.



"Who was the first black president in Mexico? [Context] → Prediction: AMG

What are the names of all the stores located in Khanewal District? [Context] → Prediction: KUF

What is motto of the state whose official symbol is cranberry? [Context] → Prediction: "

Figure 1: Example of training instruction constructed from synthetic TeaBReaC dataset where demonstrations share analogical reasoning chain. In Concept-aware Training (CoAT), we construct such examples to train in-context learners to utilise latent reasoning concepts whenever available in demonstrations.

Modern in-context learners may perform ICL with quality comparable to task-specialized models (Zhao et al., 2023; Štefánik et al., 2023). However, it remains unclear why some LMs are able of ICL in such quality while others are not; Initial work introducing GPT3 (Brown et al., 2020) followed by Thoppilan et al. (2022); Chowdhery et al. (2022); *inter alia* explains ICL as an emergent consequence of models' scale. But more recent LMs (Sanh et al., 2022; Wang et al., 2022; Wei et al., 2021; Ouyang et al., 2022) are based on 10 to 100 times smaller models and reach comparable ICL quality, instead attributing the ICL ability to a vast volume and diversity of pre-training tasks and instructions.

On the contrary, theoretical studies uncover different determinants of ICL quality than the model or data scale, relating ICL to specific data qualities, such as the occurrence of cases that can *not* be explained by mere statistical co-occurrence of tokens. Notably, Xie et al. (2022) specify this as the occurrence of training exemplars that can *only* be resolved by identifying **latent concepts**, i.e. underlying *functional* relations that *explain* the correct prediction. In this and other work surveyed in Sec-

[*]Corresponding author: stefanik.m@mail.muni.cz

Findings of the Association for Computational Linguistics: ACL 2024, pages 12335–12352
August 11-16, 2024 ©2024 Association for Computational Linguistics

tion 2, authors prove that ICL can also emerge with *both* small data *and* small models.

Our work explores the *practical* potential of concept-dependent data on the quality and robustness of in-context learning. In Section 3, we propose and implement a data construction framework that *encourages* the occurrence of concept dependencies in training data, and hence, *requires* models to learn to utilise latent concepts that explain these irregularities (Fig. 1). We call this framework **Concept-aware Training** (**CoAT**).

In Sections 4, we explore the impact of CoAT in controlled settings. We find that (i) it is possible to train language models for in-context learning of *unseen* latent concepts and (ii) that such concept-aware in-context learning *is more robust* to the functional deficiencies of existing in-context learners. Finally, on a set of over 70 tasks of SuperGLUE and Natural-Instructions, we find that CoAT can also improve practical in-context learning performance over traditional instruction tuning approach; in many cases, CoAT enables ICL of otherwise not learnable tasks, and with only two training tasks reaches ICL performance *comparable* to in-context learners of similar or larger size trained on massive collections of over 1,600 tasks.

2 Background

Methods for training in-context learners In-context learning ability, including few-shot ICL, was first uncovered in GPT3 (Brown et al., 2020) trained unsupervisedly for causal language modelling. With no other substantial differences to previous GPT models, the emergence of ICL was attributed to GPT3's *scale*, having grown to over 170-billion parameters since GPT2 (≈800M params).

Not long after, a pivotal work of Schick and Schütze (2020) on a Pattern-exploiting training (PET) has shown that even much smaller (110M) models like BERT (Devlin et al., 2019) can be fine-tuned using self-training in a similarly small data regime, first disputing the assumption on the necessity of the scale in rapidly learning new tasks.

A line of generation models further undermined the assumption of the size conditioning of ICL. Among the first, Min et al. (2022a) fine-tune smaller models (<1B parameters) on a large mixture of tasks in the few-shot instructional format and find that such models can perform previously unseen tasks. Following approaches (Sanh et al., 2022; Wang et al., 2022) also train smaller models for instruction following on large mixtures of tasks, assuming that the model's ability to in-context learn an unseen task emerges from a large diversity of instructions and task types. A recently popularised reinforcement learning approach of InstructGPT (Ouyang et al., 2022) also presents an adaptation of an instruction-following objective, training on mixtures of instructions with automatic feedback.

Recently, the instruction following was extended by joint training on programming code generation (Chen et al., 2021) and by Chain-of-Thought (CoT) targets (Wei et al., 2022), where the model is trained to respond with a sequence of natural-language steps deducing the answer (Zhao et al., 2023; Kadlčík et al., 2023). These extensions were empirically shown to enhance ICL ability (Fu and Khot, 2022) and were adopted by Flan models (Chung et al., 2022).

Analyses of ICL Recent studies shed some light on the functioning of ICL in LMs through controlled experimentation, finding that the LMs' decision-making in ICL does not align with humans. Notably, Lu et al. (2022) report on the sensitivity of LMs to the specific formulation of the instructions in the prompt, while Liu et al. (2022) measures sensitivity to the ordering of in-context demonstrations. Further, we find that LMs perform ICL comparably well when the labels of the demonstrations are randomly shuffled (Min et al., 2022b) or when the presented CoT sequences do not make sense (Wang et al., 2023). We note that such behaviours *differ* from learning a *functional* relation from demonstrations that we expect from in-context learners (Li et al., 2023b) and can be *exploited* to lead models to incorrect predictions.

Nevertheless, other studies report that under the right conditions, LMs *are* able to learn functional relationships *solely* from the input prompt; For instance, Akyürek et al. (2023); Li et al. (2023c) show that Transformers can be trained to accurately learn regression functions solely from the prompt.

Xie et al. (2022) identify the key covariate of ICL quality in the occurrence of training examples where correct predictions are conditioned by *latent concepts*. Consider a pre-training example 'Albert Einstein was [MASK]'; The correct prediction for [MASK] *can* be resolved if the model can *extract* and *apply* a latent reasoning concept from context, such as that the context exhibits a concept of *nationalities* and hence, [MASK] best substitutes 'German'. Such concept dependencies occur in

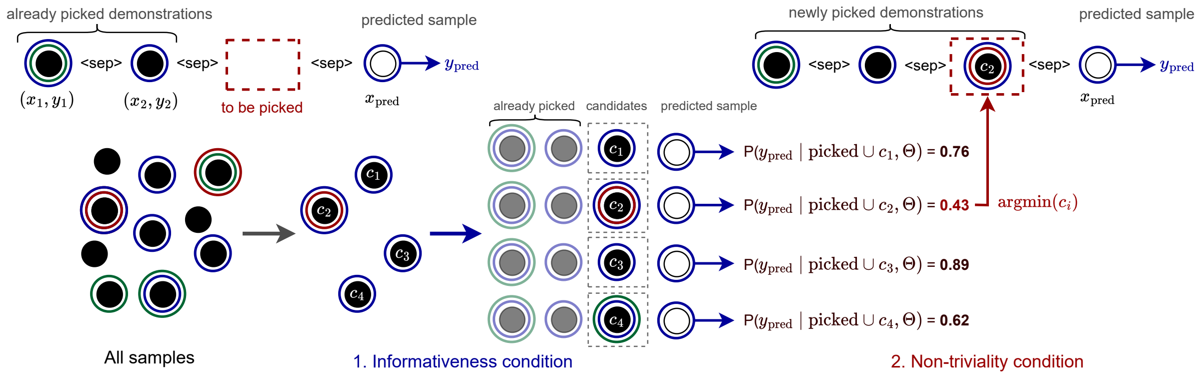

already picked demonstrations
<sep>
<sep>
<sep>
(x_1, y_1)
(x_2, y_2)
to be picked
predicted sample
x_{pred}
y_{pred}
All samples
c_1
c_2
c_3
c_4
1. Informativeness condition
already picked
candidates
predicted sample
P(y_{pred} | picked ∪ c_1, Θ) = 0.76
P(y_{pred} | picked ∪ c_2, Θ) = 0.43
P(y_{pred} | picked ∪ c_3, Θ) = 0.89
P(y_{pred} | picked ∪ c_4, Θ) = 0.62
argmin(c_i)
2. Non-triviality condition
newly picked demonstrations
<sep>
<sep>
<sep>
predicted sample
x_{pred}
y_{pred}

Figure 2: **Demonstrations selection in Concept-aware training (CoAT):** From all samples of the training dataset, we first (1) filter out ones *sharing* a specific reasoning concept ○ with predicted sample ($x_{\text{pred}}, y_{\text{pred}}$). From this subset, we (2) iteratively pick the candidate demonstration(s) c_i such that the trained model Θ's probability of generating the correct prediction y_{pred} if we pick c_i among demonstrations is *minimal*.

language sparsely but naturally, allowing the emergence of a certain ICL quality in LMs (Wies et al., 2023). Later work attributes ICL to similar data properties labelled under the terms of statistical *burstiness* (Chan et al., 2022) or *compositionality* (Hahn and Goyal, 2023).

Our work builds upon this theory, but compared to previous studies limited to in-silico experiments, we elevate the idea of concept-aware training to real-world settings, with publicly available datasets and pre-trained language models. We are first to measure the impact of concept-aware data construction in *extrinsic* evaluation over 70 diverse tasks and show its potential to substantially enhance data efficiency and robustness in training in-context learners, compared to previous work using magnitudes of more data and compute.

3 Concept-Aware Training

Aiming to create language models able to learn a new latent reasoning concept in-context, we propose a **Concept-Aware Training** (CoAT) as an instruction-tuning framework specifying **conditions for a selection of few-shot demonstrations** for the training instructions (Figure 2).

We assume the format of training prompts widely used in the previous work training in-context few-shot learners, constructing training instructions from k demonstrations consisting of the input texts x with labels y followed by the predicted sample's input text x_{pred}:

$$[x_1, y_1, \langle sep \rangle, \ldots, x_k, y_k, \langle sep \rangle, x_{\text{pred}}] \rightarrow y_{\text{pred}}$$

In this setting, CoAT proposes to filter in-context demonstrations sequentially by two conditions. The main condition, denoted as **informativeness condition**, assures to pick demonstrations exhibiting a specific *reasoning concept* C that is *shared* between a picked demonstration (x_i, y_i) and the predicted example $(x_{\text{pred}}, y_{\text{pred}})$, thus picking only the demonstrations whose reasoning pattern is *informative* for the correct prediction. Such settings make it beneficial for the trained model to learn to *extract* and *apply* concepts presented in demonstrations.

However, as the sole *informativeness* condition may easily pick demonstrations very similar or identical to the predicted sample, we propose a second, **non-triviality condition**. This condition chooses from the informative demonstrations the ones with which it is 'difficult' for the model to respond correctly. This condition avoids the occurrence of in-context demonstrations *identical* to the predicted sample and may also increase the heterogeneity of different concepts that co-occur among the demonstrations, avoiding the over-reliance on the presence of a small set of specific concepts in small-data settings.

We note that a body of previous work proposes better ways for picking in-context demonstrations during the *inference* (without updating the model) (Li et al., 2023a; Gupta et al., 2023; Luo et al., 2024). While some of these strategies are applicable in picking informative demonstrations in CoAT, note that this line of work is complementary to ours in assuming an existing in-context learner. More importantly, our motivation is substantially different; CoAT uses demonstrations of instruction training as a vehicle for creating training cases conditioned on latent concepts. This idea is not restricted to instruction tuning, but we note that instruction tuning presents an opportunity to implement it easily.

What constitutes a concept applicable in CoAT? We broadly define the term *concept* as an arbitrary functional relation of input and prediction (Xie et al., 2022) that holds robustly for *any* sample of a given task. Hypothetically, once the model learns to *model* a specific concept perfectly, it will never produce a prediction that violates this concept (Štefánik and Kadlčík, 2023). In practice, attempts to clearly present useful concepts to the model can be obstructed by other covariates (Mikula et al., 2024), such as frequent predictive co-occurrences of tokens. Thus, we propose to pick CoAT's concepts among features that are unlikely to be substitutable by non-robust covariates, presenting as best candidates to be features conditioned on a deep or holistic decomposition of the input.

Note that the goal of CoAT is *not* to represent chosen training concepts in the model's weights *explicitly* but to improve the model's ability to *extract* and *apply* available concepts from demonstrations. Towards this goal, CoAT fundamentally assumes that the ability to extract and apply one concept transfers to *other* concepts beyond the training distribution. We verify this ability later in Section 4.4.

3.1 Proposed Implementation

In our experiments, we implement the proposed CoAT framework in two training stages: First, we train LM on a scalable synthetic QA dataset, which, contrary to traditional QA datasets, contains annotations of reasoning concepts. Second, we refresh the LM's ability to work with natural language prompts by further training on a QA dataset with only natural language inputs. Hence, contrary to previous work utilising massive multitask training, in total, we only use *two* QA datasets.

Informativeness condition We find a large collection of annotated reasoning concepts in a TeaBReaC dataset of Trivedi et al. (2022), containing more than 900 unique explanations over a large set of *synthetic* QA contexts. Each TeaBReaC's explanation maps a natural question to the answer span through a sequence of declarative *reasoning steps*, such as "select→group→project". Within CoAT, we use these explanations as the shared concepts C (Fig. 1); In the training prompts, all demonstrations exhibit the same reasoning chain as the predicted sample.

To restore the model's ability to work with a natural language, in the second step, we fit the resulting model to *natural* inputs by further fine-tuning on AdversarialQA dataset (Bartolo et al., 2021); As the annotations of reasoning concepts in general QA datasets are scarce, in this case, we naively use the initial word of the question ("Who", "Where", ...) as the shared concept, aware that such-grouped samples are not always mutually informative.

Non-triviality condition In both training stages, we implement the *non-triviality condition* in the following steps. (i) We select a random *subset* of 20 samples that passed the *informativeness* condition (denoted X_{info}). (ii) From X_{info}, we iteratively *pick* a sequence of $i \in 1..k$ demonstrations (with $k : 2 \leq k \leq 8$) as follows:

1. For each sample $(x_j, y_j) \in X_{\text{info}}$, we use the training model to compute a *likelihood* of generating the correct prediction y_{pred} if a given sample (x_j, y_j) is included among demonstrations. Whenever y_{pred} contains more than one token, we compute the likelihood as the *average* of the likelihoods of *all* y_{pred}'s tokens in the teacher-forced generation.

2. In each step i, we pick among the demonstrations a sample with which the likelihood of generating correct prediction is *minimal*.

An overview of this process is depicted in Figure 2, with a schematic example of a training prompt in Figure 1. Full training prompts that our implementation of CoAT constructs in training on each dataset can be found in Table 2 in the Appendix.

4 Experiments

Our experiments provide empirical evidence towards answering three research questions (**RQs**):

1. **Can we improve models' ability to *benefit* from new reasoning concepts in-context?**

2. **Are the concept-aware in-context learners more *robust* to known functional artifacts?**

3. **Can concept-aware in-context learning also improve performance in *real-world* tasks?**

The first two RQs validate our assumptions on concept-aware training: that (1) the implementation of CoAT indeed *improves* models' utilisation of both seen and out-of-distribution concepts from demonstrations, and that (2) such an ability *can* make the in-context learning of a CoAT-trained language model more robust to artefacts revealed in previous in-context learners (Wei et al., 2023). Finally, in (3), we assess whether the enhanced models' ability to rely more on latent concepts can also improve the practical quality of low-resource in-context learning.

4.1 Training and Evaluation

To maximise comparability with the previous work, we fine-tune our models from mT5 pre-trained models of Xue et al. (2021). In both training stages (Sec. 3.1), we fine-tune all model parameters in a teacher-forced next-token prediction until convergence of evaluation loss.[*] We further detail the training parameters in Appendix A.

In all experiments, we construct evaluation prompts from $k = 3$ demonstrations chosen consistently for all models, with prompts including the options for expected labels. We complement all our evaluations with confidence intervals from the bootstrapped evaluation (population $n = 100$, repeats $r = 200$). We specify evaluation setup separately for each experiment (§4.4–4.6) with further details and examples in Appendix B.

4.2 Baselines

We assess the impact CoAT's main design choices against two baselines, allowing us to measure the impact of both its data construction conditions.

Random demonstrations selection (Tk-random) We evaluate the impact of all CoAT's components against a baseline trained in identical settings but picking the in-context demonstrations *randomly* with uniform probability over the whole training set. This baseline reproduces the methodology of a majority of the referenced work on instruction tuning, including Tk-Instruct (Wang et al., 2022) and Flan (Chung et al., 2022). Apart from the demonstration selection, all other settings, including training data, remain identical (§4.1) to assure comparability with CoAT models.

Demonstrations passing only *informativeness* condition (Tk-info) In this baseline, we perform ablation of CoAT's *non-triviality* condition (Sec. 3) by picking the demonstrations passing *only* the *informativeness* condition. Hence, such-picked demonstrations in the training instructions are informative for the prediction but can exhibit cases where some of the demonstrations are similar or even identical to the predicted sample, making it trivial for the model to perform correct prediction. All other training settings are unchanged (§4.1).

4.3 Other evaluated models

We also evaluate three recent in-context learners for which we can assess which datasets were used in their training mix: (1) **T0** of Sanh et al. (2022) trained on a mixture of 35 datasets of different tasks in zero-shot settings, mostly of QA type, mapped into a self-containing human-understandable interaction format; (2) **Tk-Instruct** of Wang et al. (2022) pre-trained in a few-shot format similar to ours, on a mixture of 1,616 diverse tasks, and (3) **Flan** models of Chung et al. (2022) that further extend data settings of Tk-Instruct to a total of 1,836 tasks, including chain-of-thought labels, i.e. a step-by-step reasoning chain mapping input prompt to a label.

All these models are based on the same pre-trained model (T5), making the results comparable to the level of fine-tuning methodology. Tk-Instruct and Flan use the data construction reproduced in our Tk-random baseline, but applied in vastly larger data settings.

4.4 CoAT's ability to improve models utilisation of latent concepts (RQ1)

We pose that if the model can utilize a new reasoning concept C from demonstrations, it will be able to *improve* the prediction in cases where the demonstrations use the same C as the predicted sample. Thus, to evaluate if training with CoAT improves models' utilisation of concepts, we evaluate models' performance in a few-shot setting where we ensure that the demonstrations *share* a specific latent concept with the predicted sample. Then, we quantify models' ability to *improve* from the concept by computing the *difference* in accuracy between such concept-sharing evaluation and conventional evaluation using *randomly* chosen demonstrations.

We perform the first analysis on TeaBReaC with annotated *reasoning chains* as concepts C shared between demonstrations and predicted sample (Fig. 1), but to evaluate generalization to *unseen* concepts, we filter out all samples with reasoning chains that were present in training. This results in 316 evaluation scenarios presenting models with 14 previously unseen reasoning patterns. In this setting, we compare the concept-improving ability of CoAT models with Tk-random baseline.

[*] Implementations of CoAT training and concept-learning evaluations are on https://github.com/MIR-MU/CoAT

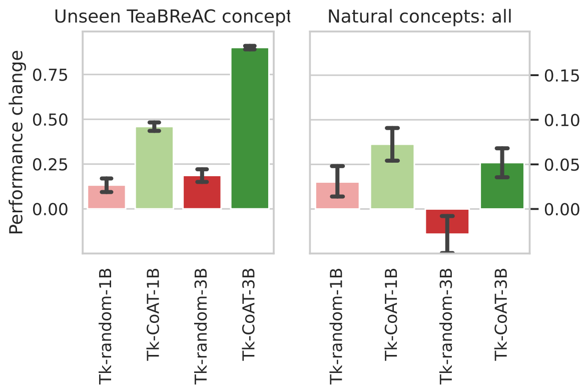


Figure 3: **In-context learning of new concepts**: Relative change of performance of models when presented with demonstrations exhibiting a reasoning concept informative for prediction. Evaluation with (left) synthetic TeaBReaC samples, and (right) diverse concepts of *natural* datasets (§4.4).

The important limitation of evaluation with TeaBReaC's concepts is that it remains unclear whether evaluation with *synthetic* samples is representative for concept learning in a *natural* language. Hence, we also perform this analysis with samples and concepts of natural-language tasks.

Previous work of Štefánik and Kadlčík (2023) evaluated ICL ability over four different functional concepts, all extracted from *explanations* of natural-language datasets. We adopt the concepts of this work and evaluate models for in-context learning of the following concepts: (i) *reasoning logic* of NLI samples of GLUE-Diagnostic dataset (Wang et al., 2018), (ii) *entity relations* annotated in human explanations (Inoue et al., 2020) in the HotpotQA dataset (Yang et al., 2018), (iii) *functional operations* annotated in general elementary-grade tests of OpenBookQA (Mihaylov et al., 2018), and (iv) shared *facts* in science exams of WorldTree dataset (Jansen et al., 2018; Xie et al., 2020). Examples of prompts with concept-sharing demonstrations for these datasets are shown in Table 3.

Identically to the case of synthetic concepts, we evaluate the ability of CoAT models to benefit from these concepts when exhibited in demonstrations and compare to uncontrolled demonstrations' selection (Tk-random) used in previous work.

Results

Concept-aware training improves the ability to benefit from unseen concepts Figure 3 evaluates models' ability to *improve* from presented concepts as the relative difference in performance between random and concept-sharing demonstration selection. First, evaluation with unseen TeaBReaC concepts (left) assesses models' ability to extrapolate the utilisation of latent concepts to 14 previously unseen reasoning chains. Both CoAT and random-demonstration models (§4.2) can improve from concepts presented in demonstrations. However, the improvement of CoAT-trained models is significantly larger and exceeds gains of Tk-random by 2-fold and 4-fold with the smaller and larger model, respectively. This comparison verifies that CoAT's data construction really improves our targeted skill of better utilizing concepts of demonstrations.

CoAT applied with synthetic data also improves the use of *natural* concepts Evaluation of improvements on selected natural concepts (Figure 3; right) shows that concept-learning ability obtained with synthetic TeaBReaC concepts *transfers* to natural-language settings, as the CoAT-trained models can benefit from concepts significantly *more* than models trained without concept-aware data construction (Tk-random).

Despite that, evaluations over the individual reasoning concepts (Figure 7 in Appendix C.3) reveal that even CoAT models can not benefit robustly from *all* concepts. Nevertheless, we note that in the cases where CoAT models do not improve, also *none* of the baselines benefit from presented concepts. This might be attributed to several reasons: (i) the presented concepts are not really *informative* for prediction, (ii) our training data allowed the models to *memorize* relevant knowledge and, hence, do not *need* (and *benefit from*) the concepts' exposition, or (iii) our training concepts were simply not sufficient to generalize over these new concepts.

4.5 Robustness of concept-aware in-context learners (RQ2)

As we overviewed in Section 2, other work reports functional deficiencies of previous in-context learners, including surprising insensitivity of in-context learners to the assigned demonstrations' labels (Min et al., 2022b). Wei et al. (2023) attribute this to models' over-reliance on the *semantic priors* obtained in pre-training, which may override learning of the *functional* concepts. Such behaviour is defective, because the ability to learn functional relations is necessary for robust and interpretable in-context learning of truly unseen tasks.

To evaluate the impact of concept-aware training on models' reliance on their semantic priors,

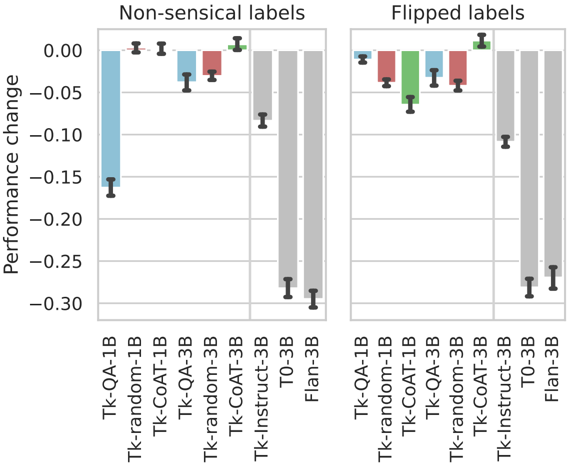


Figure 4: **Models' reliance on semantic priors**: Relative change of models' performance when we **(left) replace** labels with 'non-sensical' tokens with no correspondence to the semantics of the task, such as '*foo*', '*bar*', etc.; and **(right) flip** the original labels, so that e.g. '*negative*' label corresponds to a positive-sentiment sample. CoAT models can in-context learn the input-output mapping similarly well with non-sensical labels and rely on the labels' semantics significantly less than previous in-context learners (in grey).

we follow the setup of Wei et al. (2023) and assess reliance on *labels*' semantics in a standard few-shot evaluation (§4.1), with one of the two modifications; (i) We change the labels to tokens with *irrelevant* meaning for the prediction task, such as 'Foo', 'Bar' etc. (ii) We *shuffle* the labels so that semantically incorrect labels are assigned in the demonstrations, but the input-label mapping remains consistent. In both settings, the task's functional relation can still be recovered from demonstrations, but the sole reliance on semantics will either not help or will mislead the model.

In this setting, we evaluate three model types: (i) CoAT-trained models, (ii) models with uncontrolled data construction (Tk-random & previous work), and (iii) models with uncontrolled data construction, but fine-tuned *only* on a *natural* QA dataset (denoted Tk-QA). We perform the evaluation over 8 SuperGLUE tasks with discrete labels.

Results Figure 4 shows the results. Evaluation with non-sensical labels (left) reveals that all models pre-trained on a synthetic TeaBReaC dataset (Tk-random, and Tk-CoAT) are more robust to the labels' semantics than our natural-language baseline (Tk-QA). However, a comparison of Tk-random and Tk-CoAT suggests that Tk-CoAT's preference for learning functional relations is a composition of *both* using a synthetic dataset in pre-training *and* CoAT's data construction.

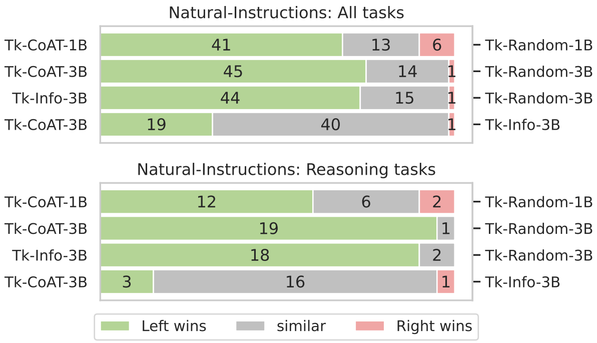


Figure 5: **Effectiveness of Concept-aware training: Natural-Instructions:** Win rate of models utilising Concept-aware training (CoAT; §3) and traditional instruction tuning (Tk-Random; §4.2) evaluated on (top) *all* and (bottom) *reasoning* tasks of Natural-Instructions collection. Values indicate the number of tasks where the referenced model reaches significantly higher accuracy than the other. For the *similar* tasks, the difference in models' performance is not statistically significant.

A comparison to previous models reveals that all multitask models experience substantially larger decay in performance than our models. We suspect this feature could be a *bias* specific to massive multi-task learning emerging when label semantics can *explain* a large portion of training data. This result is consistent with Wei et al. (2023), but contrary to their conclusions, we show that ICL robust to semantic distractions does *not* emerge *exclusively* with very large (≥ 100B) model scale.

Nevertheless, we note that the smaller CoAT model still relies on labels' semantics when recognizable (*Flipped labels*; Fig. 4 right), less than previous models, but comparable to our baselines.

4.6 Practical effectiveness of concept-aware in-context learners (RQ3)

Finally, we assess the practical quality of concept-aware ICL on previously unseen tasks in a simulated low-resource application with only three *randomly-chosen* demonstrations. We evaluate on samples from two collections of tasks: (i) SuperGLUE (Wang et al., 2019) consisting of 10 tasks requiring a variety of reasoning skills, and (ii) a test split of Natural-Instructions (Wang et al., 2022) from which we pick 60 extractive tasks. For SuperGLUE tasks, we verbalize both the demonstrations and predicted sample using all available templates within PromptSource library (Bach et al., 2022) and report results for the best-performing template

	AxG	Ax-b	WSC	CB	RTE	WiC	ReCoRD	BoolQ	COPA	MultiRC
Tk-random-1B	49.4±5.2	43.6±4.8	52.7±5.1	21.8±3.9	29.3±4.6	18.0±4.0	15.3±3.8	34.0±5.0	74.7±3.4	5.1±2.4
Tk-random-3B	50.2±5.4	57.5±4.8	52.0±5.5	47.8±5.1	48.9±4.8	50.1±4.4	16.3±7.3	62.8±4.6	75.5±2.8	2.1±1.5
Tk-info-1B	50.0±4.2	42.6±5.7	52.0±4.3	47.2±3.9	49.2±4.8	53.2±4.5	15.5±4.0	19.6±2.3	61.5±2.3	3.2±1.2
Tk-info-3B	50.8±4.6	57.2±4.9	53.5±4.8	47.3±5.4	54.7±4.9	53.6±4.7	22.6±4.5	64.4±4.8	76.3±3.0	2.7±2.1
Tk-CoAT-1B	50.4±5.3	52.7±4.6	53.6±5.2	46.9±4.9	53.7±4.9	53.5±5.3	17.0±3.5	63.8±5.4	76.1±3.2	11.4±2.6
Tk-CoAT-3B	57.9±4.9	57.2±4.8	53.6±4.5	60.4±4.8	52.0±5.4	56.9±5.0	23.1±3.8	63.6±4.3	81.3±3.3	56.9±3.6

Table 1: **Effectiveness of concept-aware training: SuperGLUE:** ROUGE-L scores of ICL models evaluated in few-shot setting on SuperGLUE tasks (Wang et al., 2019), trained using (i) *random* demonstrations sampling used in previous work, (ii) *informative* demonstrations sampling (§4.2) and (iii) *informative+non-trivial* sampling (CoAT; §3). Underlined are the best results per each task and model size. See Table 5 for a comparison to previous models.

for each model. For Natural-Instructions tasks, we prefix the demonstrations with the instruction provided with each task. To maximise evaluation reliability over all models, we analyse the error cases and choose to report the results in ROUGE-L for SuperGLUE, and in a standard accuracy for Natural-Instructions. We specify the metrics selection analysis and other evaluation details in Appendix B, with prompt examples in Table 4.

As a primary reference point, we again compare the results of CoAT-trained models to Tk-random, where we can make sure that all other training configurations except for the data construction method are identical. Further, we compare to Tk-info (without *Non-triviality* condition; §4.2) to also evaluate the importance of the *non-triviality* condition. Finally, to provide additional context to our results, we also compare the performance of CoAT-trained models to previous in-context learners (§4.3).

Results Figure 5 compares the accuracy of CoAT-trained models to our baselines: (i) without systematic demonstrations selection (Tk-Random) and (ii) without the *non-triviality* condition (Tk-Info), over 60 tasks of NaturalInstructions collection. In comparison to Tk-Random, CoAT models reach significantly higher accuracy on 41 and 45 of 60 tasks, with comparable performance on a majority (13 and 14) of other tasks. The difference is further magnified on *reasoning* tasks, which we argue might better evaluate models' ability to in-context learn a *functional* relation of the new task. A comparison of Tk-Info with Tk-Random shows that the performance on reasoning tasks is mainly fostered by the CoAT's *informativeness* condition, but in a full task collection, Tk-CoAT still outperforms Tk-Info in 19 out of 60 tasks. Evaluations on other task segments can be found in Appendix C.2.

In the evaluation over the tasks of SuperGLUE collection (Table 1), we additionally report the specific values of ROUGE-L that our baselines and CoAT models achieve. With a single exception, models utilising a concept-based selection of demonstrations (Tk-CoAT and Tk-Info) consistently reach higher scores than Tk-Random. Our analyses of models' predictions reveal that in 7 out of 20 evaluations, Tk-Random models fail to follow the task's instruction, consequentially responding out of valid label space. Tk-CoAT is shown to mitigate this issue in all cases except for a smaller CoAT-trained model on MultiRC. A comparison of Tk-CoAT with Tk-Info shows that *non-triviality* condition is more substantial for a smaller model, but the models of both sizes benefit similarly from the concept-sharing selection of demonstrations.

Comparison to multitask learners Figure 6 contextualizes the performance of CoAT models trained on two datasets of a single (QA) task with existing instructional models trained on massive mixtures of 35–1,836 tasks. Over all the NI tasks (Fig. 6; *top*), CoAT models outperform multitask learners on a majority of tasks in 3 of 6 competitions. CoAT models are outperformed by Flan models but perform at least *similarly* for the *majority* of the tasks in 5 out of 6 competitions. The evaluation on reasoning tasks (Fig. 6; *middle*) supports our hypothesis that CoAT particularly promotes improvements in in-context learning of new *reasoning* abilities, winning on reasoning tasks over Flan and Tk-Instruct in a comparable number of cases than the opponents.

Finally, we look at a few tasks with *unseen labels* for both Tk-Instruct and Flan models (Fig. 6; *bottom*) where multitask learners can not rely on shortcuts based on unseen tasks' labels. Here, the results of competition between CoAT with Flan models *turns over*, with CoAT models performing significantly better on 4 out of 6 tasks. While this sole segment is not big enough for robust conclusions, the results further support our claim that

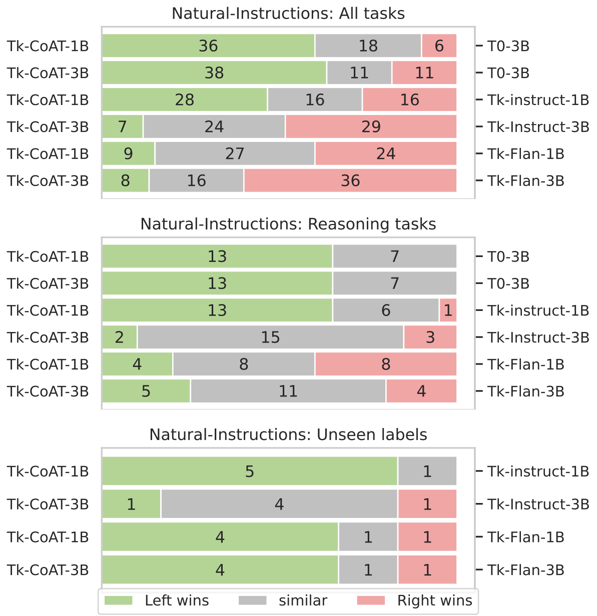


Figure 6: **Performance comparison to previous work: Natural-Instructions collection:** Win rate of CoAT models trained using two (2) tasks and previous in-context learners trained on mixtures of 35 (T0), 1,616 (Tk-Instruct) and 1,836 tasks (Tk-Flan). Values denote the number of tasks where the model reaches significantly better accuracy. Evaluations over (top) all tasks, (middle) reasoning tasks, (bottom) tasks with labels not present in the training mix of Tk-Instruct and Tk-Flan.

concept-based ICL is *more robust* to semantic distractions (RQ2).

Table 5 in Appendix C details models' scores on SuperGLUE tasks, providing further evidence on overall comparability of CoAT models to multitask learners. For instance, a comparison with Tk-Instruct reveals that CoAT's 1B and 3B models reach higher absolute scores on 3 and 5 out of the 7 Tk-Instruct's unseen tasks.

5 Conclusion

Inspired by data-centric theories on emergence of in-context learning, we propose and implement a Concept-aware Training framework for constructing training scenarios that challenge language models to learn to *utilise* latent concepts from in-context prompts. We show that language models *can* be trained to benefit from *unseen* concepts (RQ1), and that such ICL *is* more robust in learning *functional* relations of a new task from demonstrations (RQ2). Finally, in extrinsic evaluation over 70 tasks, we demonstrate the practical efficiency of concept-dependent training data, with CoAT models bringing significant improvements on 41 and 45 out of 60 Natural-Instructions tasks, or 6 and 5 of 10 SuperGLUE tasks (RQ3), while reaching a performance comparable to multitask learning requiring *magnitudes* of more data.

More broadly, our work pioneers an alternative direction for scaling the quality of in-context learning to the previously explored *model* and *data scale* axes. We wish towards inspire future work to a more proactive approach to refining training data properties so that fitting such data *necessitates* the emergence of the specific, robust abilities of the models, such as the concept-learning ability.

Specifically, future work can build upon our findings in researching ways to upscale concept-dependent data in unsupervised settings, allowing for pre-training more robust language models with a fraction of data and computing budget.

Limitations

Although our main objective is to assess the efficiency of concept-aware training, we acknowledge the limitations of our comparison to the previous work, where several aspects convolute the representative comparison of different in-context learners: (i) each of the multitask learners was trained on a different, yet massive set of tasks, making it difficult to find a broader collection that is *new* for multiple models; For this purpose, we surveyed three standard collections used for few-shot evaluation: CLUES (Mukherjee et al., 2021), RAFT (Alex et al., 2021) and FLEX (Bragg et al., 2021), but found in total only three tasks unseen by the multitask learners of previous work, all of the same type (classification). Therefore, in our evaluations, we use (a) Tk-Instruct's own evaluation set and (b) SuperGLUE, which significantly overlaps with the training tasks of previous work. (ii) many aspects make it "easier" for the model to improve, including the domain of labels or prompt format matching the training distribution (relevant to Tk-Instruct and Flan evaluated on Natural-Instructions).

Another aspect that we neglect in our experiments in favour of more in-depth analyses is the *impact of pretraining* projected into the properties of the foundation model that we use. We pick T5 as a base model to maximise comparability with previous work. While we do not identify any concrete reason to assume that CoAT would perform worse with other base models, one should note that our results do not provide any evidence in this respect.

Finally, we note that the applicability of CoAT is conditioned by the availability of the annotated *concepts* C in the training datasets, which might be difficult to obtain for natural-language datasets. Our implementation circumvents this issue by using a synthetically curated dataset. Hence, we simultaneously show that concept-aware abilities can also be obtained in the restrictive settings of synthetic-dataset pre-training, where we note that the volume and variability of the synthetic dataset can be scaled further much easier than the natural dataset(s) (Trivedi et al., 2022). Nevertheless, our experiments do not provide any empirical evidence for answering *to what extent* could further extension of synthetically-generated datasets, possibly covering even more complex concepts, *scale* to further performance gains.

Ethical Considerations & Broader Impact

The primary motivation of our work is to minimise the computing demands for the creation of accurate in-context learners by deepening our understanding of the covariates of the resulting quality. We believe that our presented method, as well as the future data-efficient methods improving our understanding of in-context learning, will enable the democratization of the creation of robust and accurate in-context learning models for both research and industry.

Finally, we note that data-efficient methods for training ICLs (as opposed to *multitask training*) might open possibilities for creating more accurate ICLs specialized to languages outside English, where training datasets are scarce. We look forward for the future work that will explore the potential of data-efficient instruction tuning specifically on the target-language datasets, creating in-context learners specially tailored for target languages outside English.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? Investigations with linear models. In *The Eleventh International Conference on Learning Representations*.

Neel Alex, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C. Jess Riedel, Emmie Hine, Carolyn Ashurst, Paul Sedille, Alexis Carlier, Michael Noetel, and Andreas Stuhlmüller. 2021. RAFT: A real-world few-shot text classification benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Xiangru Tang, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation. In *Proceedings of the 2021 Conference EMNLP*, pages 8830–8848, Online and Punta Cana, Dominican Republic. ACL.

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. FLEX: Unifying Evaluation for Few-Shot NLP. In *Advances in Neural Information Processing Systems*, volume 34, pages 15787–15800. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in NIPS*, volume 33, pages 1877–1901. Curran Associates, Inc.

Stephanie C.Y. Chan, Adam Santoro, Andrew Kyle Lampinen, Jane X Wang, Aaditya K Singh, Pierre Harvey Richemond, James McClelland, and Felix Hill. 2022. Data Distributional Properties Drive Emergent In-Context Learning in Transformers. In *Advances in Neural Information Processing Systems*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N.

Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *arXiv e-prints*, page arXiv:2210.11416.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of the 2019 Conference of the NAACL: Human Language Technologies*, pages 4171–4186, Minneapolis, USA. ACL.

Hao Fu, Yao; Peng and Tushar Khot. 2022. How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Sources. *Yao Fu's Notion*.

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based Example Selection for In-Context Learning. In *Findings of the ACL: EMNLP 2023*, pages 13924–13950, Singapore. ACL.

Michael Hahn and Navin Goyal. 2023. A Theory of Emergent In-Context Learning as Implicit Structure Induction.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 6740–6750, Online. ACL.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marek Kadlčík, Michal Štefánik, Ondrej Sotolar, and Vlastimil Martinek. 2023. Calc-X and Calcformers: Empowering Arithmetical Chain-of-Thought through Interaction with Symbolic Systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12101–12108, Singapore. ACL.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023a. Unified Demonstration Retriever for In-Context Learning. In *Proceedings of the 61st Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. ACL.

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023b. Transformers as Algorithms: Generalization and Stability in In-context Learning.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023c. Transformers as Algorithms: Generalization and and Stability in In-context Learning.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. ACL.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. ACL.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the*

60th Annual Meeting of the ACL (Volume 1: Long Papers), pages 8086–8098, Dublin, Ireland. ACL.

Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024. In-context Learning with Retrieved Demonstrations for Language Models: A Survey.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference EMNLP*, pages 2381–2391, Brussels, Belgium. ACL.

Lukáš Mikula, Michal Štefánik, Marek Petrovič, and Petr Sojka. 2024. Think Twice: Measuring the Efficiency of Eliminating Prediction Shortcuts of Question Answering Models. In *Proceedings of the 18th Conference of the European Chapter of the ACL (Volume 1: Long Papers)*, pages 2179–2193, St. Julian's, Malta. ACL.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022a. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work?

Subhabrata Mukherjee, Xiaodong Liu, Guoqing Zheng, Saghar Hosseini, Hao Cheng, Greg Yang, Christopher Meek, Ahmed Hassan Awadallah, and Jianfeng Gao. 2021. Few-Shot Learning Evaluation in Natural Language Understanding. In *NeurIPS 2021*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on NIPS*, NIPS '22, pages 27730–27744, Red Hook, NY, USA. Curran Associates Inc.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. ACL.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the ACL: Human Language Technologies*, New Orleans, Louisiana. ACL.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference.

Michal Štefánik and Marek Kadlčík. 2023. Can In-context Learners Learn a Reasoning Concept from Demonstrations? In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 107–115, Toronto, Canada. ACL.

Michal Štefánik, Marek Kadlčík, Piotr Gramacki, and Petr Sojka. 2023. Resources and Few-shot Learners for In-context Learning in Slavic Languages. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 94–105, Dubrovnik, Croatia. ACL.

Michal Štefánik, Vít Novotný, Nikola Groverová, and Petr Sojka. 2022. Adaptor: Objective-Centric Adaptation Framework for Language Models. In *Proceedings of the 60th Annual Meeting of the ACL: System Demonstrations*, pages 261–269, Dublin, Ireland. ACL.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Teaching Broad Reasoning Skills for Multi-Step QA by Generating Hard Contexts. In *Proceedings of the 2022 Conference EMNLP*, pages 6541–6566, Abu Dhabi, United Arab Emirates. ACL.

Michal Štefánik and Marek Kadlčík. 2023. Can In-context Learners Learn a Reasoning Concept from Demonstrations? In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 107–115, Toronto, Canada. ACL.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*, page 3266–3280. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proc. of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. ACL.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the 2022 Conference EMNLP*, pages 5085–5109, Abu Dhabi, United Arab Emirates. ACL.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently.

Noam Wies, Yoav Levine, and Amnon Shashua. 2023. The Learnability of In-Context Learning.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proc. of the 2020 Conf. EMNLP: System Demonstrations*, pages 38–45. ACL.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An Explanation of In-context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations*.

Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. 2020. WorldTree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5456–5473, Marseille, France. European Language Resources Association.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 483–498, Online. ACL.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference EMNLP*, pages 2369–2380, Brussels, Belgium. ACL.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. *arXiv e-prints*, page arXiv:1810.12885.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models.

A Training details

Table 2 shows a full training example for each stage of training: (1) TeaBReaC with synthetic contexts (top) and (2) AdversarialQA with natural-language contexts (bottom). In all our training setups, we fine-tune all model parameters for teacher-forced next-token prediction, conventionally used in training sequence-to-sequence language models. In the two training stages (TeaBReaC and AdversarialQA), we use a **learning rate** of $5e^{-5}$ and $2e^{-5}$, respectively. Other parameters remain identical between stages: effective **batch size** = 30 samples and **early stopping** with the patience of 2,000 updates based on evaluation loss on a standardized validation set of each dataset. We do not report the absolute values of evaluation loss as these are not directly comparable. In CoAT training, we use a random subsample of 20 informative examples as a candidate set for a selection of non-trivial demonstrations.

Other parameters of training configuration default to Training Arguments of Transformers library (Wolf et al., 2020) in version 4.19.1. For readability, we implement the relatively complex demonstrations' selection as a new objective of the Adaptor library (Štefánik et al., 2022). The picked demonstrations are encoded into a format consistent with the evaluation.

B Evaluation details

Tables 3 shows an example of an instruction for each evaluation that we perform within the concept-learning evaluation. For readability, we only shorten the examples of HotpotQA, where we omit some sources of data available for the model. In the case of TeaBReaC not shown in this table, the evaluation prompt format is the same as in training (Table 2), whereas we make sure that the reasoning chains of evaluation samples differ from the training.

SuperGLUE Evaluation format As mentioned in Section 4.1, we verbalize both the demonstrations and predicted sample using all available templates of PromptSource library (Bach et al., 2022), obtaining prompts for each demonstration prompt x_i and its label y_i in a free-text form. The prompts commonly contain the full-text match of the possible labels as options for the model.

Following the example of Wang et al. (2022), we additionally prepend the demonstrations and labels with keywords "Input" and "Prediction" and separate demonstrations with new lines. Thus, the resulting input→output pairs in evaluation take this format:

"Input: x_1 *Prediction:* y_1 *<newline>*
Input: x_2 *Prediction:* y_2 *<newline>*
Input: x_3 *Prediction:* y_3 *<newline>*
Input: x_{pred} *Prediction: "* → "y_{pred}"

where demonstrations (x_i, y_i) are picked randomly but consistently between all evaluated models.

Natural-Instructions Evaluation format In the evaluations on Natural-Instructions, we closely follow the example of Wang et al. (2022) and additionally prepend the sequence of demonstrations with an instruction provided for each task:

"<task instruction> <newline>
Input: x_1 *Prediction:* y_1 *<newline>*
Input: x_2 *Prediction:* y_2 *<newline>*
Input: x_3 *Prediction:* y_3 *<newline>*
Input: x_{pred} *Prediction: "* → "y_{pred}"

where the *<task instruction>* contains the instruction as would be given to the annotators of the evaluation task, usually spanning between 3–6 longer sentences. The demonstrations are again picked randomly but consistently between models.

Examples of evaluation prompts for both SuperGLUE and Natural-Instructions can be found in Table 4.

Evaluation metrics selection Previous work training in-context few-shot learners is not consistent in the use of evaluation metrics, and the choice usually boils down to either using the exact-match accuracy (Sanh et al., 2022; Chung et al., 2022) or ROUGE-L of Lin (2004) (Wang et al., 2022), evaluating the longest common sequence of tokens. We investigate these two options with the aim of not penalising the models for minor discrepancies in the output format (in the accuracy case) but avoiding false positive evaluations in predictions that are obviously incorrect (in the ROUGE case).

Investigation of the models' predictions reveals that the selection of the metric makes a large difference only in the case of Tk-Instruct models, where the situation differs between SuperGLUE and Natural-Instructions, likely due to the character of the evaluation prompts.

(1) On **SuperGlue**, e.g. on MultiRC task, for the evaluation prompt: "Does answer sound like a valid

Dataset	Concept	Training instruction	Target
TeaBReaC	**Exactly-matching reasoning chain** [*"select"* → *"maximum"* → *"list"* → *"maximum"* → *"sum"*]	“**Input:** how many points did the Monte Vesuvio" score in their two highest scoring matches? **Context:** scores of games of Pentagon". 99 scores of games of monte vesuvio". 67 scores of games of Pentagon". 6 scores of games of monte vesuvio". 76 scores of games of Pentagon". 37 scores of games of monte vesuvio". 56 scores of games of Pentagon". 8 scores of games of Pentagon". 90 scores of games of Pentagon". 20 Answer: **Prediction:** 143 **[2 more examples] Input:** how many points did the Bell 212 score in their two highest scoring games? **Context:** scores of games of bell 212. 90 scores of games of S-50. 54 scores of games of bell 212. 41 scores of games of bell 212. 36 scores of games of S-50. 23 scores of games of bell 212. 6 scores of games of bell 212. 2 scores of games of S-50. **Prediction:** ”	“131”
AdversarialQA	**Matching question-word** “Who”	“**Input:** Who was the Speaker in 1909? **Context:** Second, Democrats have always elevated their minority floor leader to the speakership upon reclaiming majority status. Republicans have not always followed this leadership succession pattern. In 1919, for instance, Republicans bypassed James R. Mann, R-IL, who had been minority leader for eight years, and elected Frederick Gillett, R-MA, to be Speaker. Mann "had angered many Republicans by objecting to their private bills on the floor;" also he was a protégé of autocratic Speaker Joseph Cannon, R-IL (1903–1911), and many Members "suspected that he would try to re-centralize power in his hands if elected Speaker." More recently, although Robert H. Michel was the Minority Leader in 1994 when the Republicans regained control of the House in the 1994 midterm elections, he had already announced his retirement and had little or no involvement in the campaign, including the Contract with America which was unveiled six weeks before voting day. **Prediction:** Joseph Cannon, R-IL. **[2 more examples] Input:** Who created the legal system still in use in Florida? **Context:** As a result of these initiatives northeastern Florida prospered economically in a way it never did under Spanish rule. Furthermore, the British governors were directed to call general assemblies as soon as possible in order to make laws for the Floridas and in the meantime they were, with the advice of councils, to establish courts. This would be the first introduction of much of the English-derived legal system which Florida still has today including trial by jury, habeas corpus and county-based government. Neither East Florida nor West Florida would send any representatives to Philadelphia to draft the Declaration of Independence. Florida would remain a Loyalist stronghold for the duration of the American Revolution. **Prediction:** ”	“British”

Table 2: Examples of **training instructions** with expected outputs, for both our datasets applied in training. Note that the shared reasoning concept is not a part of the model’s input.

answer to the question: question", Tk-Instruct-3B in our evaluation predicts "Yes." or "Yes it is" (instead of "Yes"), or "No not at all" (instead of "No"), likely due to the resemblance with the format of training outputs. As we do not wish to penalize these cases, we use ROUGE-L over all SuperGLUE evaluations.

(2) In **Natural-Instructions** evaluation, we find that Tk-Instruct often predicts longer extracts from the input prompt. This is problematic with ROUGE-L in the cases where the extract contains *all* possible answers, such as in the Tk-Instruct-1B’s prediction: “yes or no” to the prompt whose instruction ends with “Please answer in the form of yes or no.”. As we encounter this behaviour in a large portion of Natural-Instructions tasks, we evaluate all models on Natural-Instructions for exact-match accuracy after the normalization of the casing and the removal of non-alphabetic symbols. To make sure that the model is presented with the exact-matching answer option, we exclude from evaluation the tasks where the correct answer is not presented in the task’s instruction. The reference to the list of Natural-Instructions evaluation tasks can be found in Appendix C.4.

For the reported evaluations of the Reasoning tasks, we pick from the list of evaluation tasks the ones concerned with the reasoning task by simply matching the tasks with ‘reasoning’ in their name, resulting in the collection of 20 evaluation tasks.

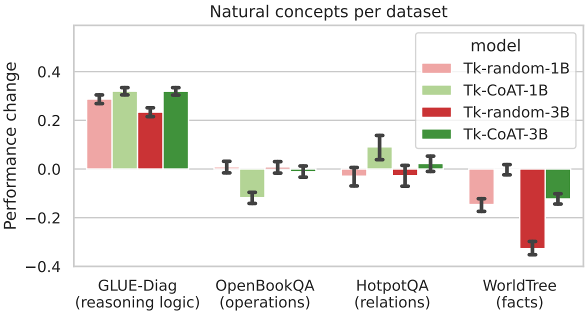


Figure 7: **In-context learning of natural concepts for each dataset**: While CoAT improves the ability to benefit from reasoning concepts on average (Fig. 3), per-concept evaluation reveals that this ability is not consistently robust.

C Further evaluations

C.1 SuperGLUE evaluations of other models

Table 5 compares the performance over the tasks of SuperGLUE collection (Wang et al., 2019) for CoAT models trained on two tasks of the same (QA) type with in-context learners trained on 35–1,836 tasks of the comparable size. Despite the

Dataset	Concept	Model instruction	Expected output
GLUE NLI Diag.	Double negation	“**Input:** I will say that she stole my money. Question: I won’t say that she didn’t steal my money. True, False, or Neither? **Prediction:** Neither **Input:** I won’t say that she didn’t steal my money. Question: I will say that she stole my money. True, False, or Neither? **Prediction:** Neither **Input:** A rabbi is at this wedding, standing right there standing behind that tree. Question: It’s not the case that there is no rabbi at this wedding; he is right there standing behind that tree. True, False, or Neither? **Prediction:** True **Input:** Even after now finding out that it’s animal feed, I won’t ever stop being addicted to Flamin’ Hot Cheetos. Question: Even after now finding out that it’s animal feed, I will never stop being addicted to Flamin’ Hot Cheetos. True, False, or Neither? **Prediction:** ”	“True”
OpenBookQA	**Shared facts:** {"Earth is greater in mass than Mars", "gravity means gravitational pull; gravitational force; gravitational attraction", "as the force of gravity increases, the weight of objects will increase."}	“**Facts:** a decrease is a kind of change. increase means more. as mass of a planet; of a celestial body increases, the force of gravity on that planet will increase. to change means to become different. an animal is a kind of living thing. the gravitational force of a planet; of a celestial object does not change the mass of an object on that planet or celestial body. an increase is the opposite of a decrease. an astronaut is a kind of human. massive means great in mass. the Mars Rover is a kind of vehicle. a living thing is a kind of object. Earth is greater in mass than Mars. gravity means gravitational pull; gravitational energy; gravitational force; gravitational attraction. greater means higher; more in value. stay the same means not changing. a moon is a kind of celestial object; body. an increase is a kind of change. Earth is a kind of planet. as the force of gravity increases, the weight of objects will increase. less is similar to decrease. Mars is a kind of planet. **Input:** An object has a weight of 10 kg on the surface of Earth. If the same object were transported to the surface of Mars, the object would have a weight of 3.8 kg. Which best explains why the weight of the object changed when transported from Earth to Mars? (A) The density of the object is greater on Earth than it is on Mars. (B) The volume of the object is greater on Earth than it is on Mars. (C) Gravitational force is greater on Earth than it is on Mars. (D) Atmospheric pressure is less on Earth than it is on Mars. **Prediction:** Gravitational force is greater on Earth than it is on Mars **[two more examples] Input:** When astronauts walked on the Moon, they used weighted boots to help them walk due to the lower gravitational pull. What difference between Earth and the Moon accounts for the difference in gravity? (A) density (B) diameter (C) mass (D) volume. **Prediction:** ”	“mass”
HotpotQA	**Shared relation in reasoning**: “X is a genus”	“**Input:** Are Broughtonia and Laeliocattleya both orchids? Hint: use the information from the paragraphs below to answer the question. Otaara, abbreviated Otr. in the horticultural trade, is an intergeneric hybrid of orchids, with "Brassavola", "Broughtonia", "Cattleya", "Laelia" and "Sophronitis" as parent genera. Paracaleana commonly known as duck orchids, is a genus of flowering plants in the orchid family, Orchidaceae that is found in Australia and New Zealand. Duck orchids have a single leaf and one or a few, dull-coloured, inconspicuous flowers. (...) **Prediction:** yes **[two more examples] Input:** Are both Parodia and Thalictrum flowering plants? Hint: use the information from the paragraphs below to answer the question. - Thalictrum ( ) is a genus of 120-200 species of herbaceous perennial flowering plants in the Ranunculaceae (buttercup) family native mostly to temperate regions. Meadow-rue is a common name for plants in this genus. - Parodia is a genus of flowering plants in the cactus family Cactaceae, native to the uplands of Argentina, Peru, Bolivia, Brazil, Colombia and Uruguay. This genus has about 50 species, many of which have been transferred from "Eriocactus", "Notocactus" and "Wigginsia". They range from small globose plants to 1 m tall columnar cacti. All are deeply ribbed and spiny, with single flowers at or near the crown. Some species produce offsets at the base. They are popular in cultivation, but must be grown indoors where temperatures fall below 10 degrees. **Prediction:** ”	“yes”
WorldTree	**Relation of objects**: "generate"	“**Input:** Despite what some think, instead around themselves, our planet spins around... Choices: pluto, the moon, the milky way, the sun. **Prediction:** the sun **Input:** In a single year, a giant globe will do this to a giant star. Choices: fight, burn, circle, explode. **Prediction:** circle **Input:** The earth revolves around... Choices: a heat source, the Milky Way, a neighboring planet, the moon. **Prediction:** a heat source **Input:** the central object of our solar system is also... Choices: the smallest object in the solar system, the coldest heavenly body, the farthest star from us, the closest star from us. **Prediction:** ”	“the closest star from us”

Table 3: Examples of **evaluation instructions** with expected outputs, for each dataset used in evaluation of in-context learning of new concepts (RQ1). Note that the demonstrations within the instructions share the annotated *Concept* with the following *predicted sample*.

Dataset	Concept	Model instruction	Expected output
SuperGLUE	-	"**Input**: The soldiers were concealed in the brush. Select the most plausible cause: - They were armed with rifles. - They wore camouflage uniforms. **Prediction**: They wore camouflage uniforms. **Input**: The print on the brochure was tiny. Select the most plausible effect: - The man put his glasses on. - The man retrieved a pen from his pocket. **Prediction**: The man put his glasses on. **Input**: I excused myself from the group. Select the most plausible cause: - I turned off my phone. - My phone rang. **Prediction**: My phone rang. **Input**: My body cast a shadow over the grass. Select the most plausible cause: - The sun was rising. - The grass was cut. **Prediction**:"	"The sun was rising."
Natural-Instructions	-	"Indicate with ‘Yes‘ if the given question involves the provided reasoning ‘Category‘. Indicate with ‘No‘, otherwise. We define five categories of temporal reasoning. First: "event duration" which is defined as the understanding of how long events last. For example, "brushing teeth", usually takes few minutes. Second: "transient v. stationary" events. This category is based on the understanding of whether an event will change over time or not. For example, the sentence "he was born in the U.S." contains a stationary event since it will last forever; however, "he is hungry" contains a transient event since it will remain true for a short period of time. Third: "event ordering" which is the understanding of how events are usually ordered in nature. For example, "earning money" usually comes before "spending money". The fourth one is "absolute timepoint". This category deals with the understanding of when events usually happen. For example, "going to school" usually happens during the day (not at 2 A.M). The last category is "frequency" which refers to how often an event is likely to be repeated. For example, "taking showers" typically occurs 5 times a week, "going to Saturday market" usually happens every few weeks/months, etc. **Input:** Sentence: Jack played basketball after school, after which he was very tired. Question: How long did Jack play basketball? Category: Event Duration. **Prediction:** Yes **Input:** Sentence: He was born in China, so he went to the Embassy to apply for a U.S. Visa. Question: How often does he apply a Visa? Category: Frequency. **Prediction:** Yes **Input:** Sentence: Jack played basketball after school, after which he was very tired. Question: Was Jack still tired the next day? Category: Event Duration. **Prediction:** No **Input:** Sentence: It refers to a woman who is dangerously attractive, and lures men to their downfall with her sexual attractiveness. Question: How long does it take to lure men to their downfall? Category: Event Duration. **Prediction:** "	"Yes"

Table 4: Examples of **evaluation instructions** with expected outputs, for selected tasks of **SuperGLUE** and **Natural-Instructions** (RQ3). Displayed samples are from CoPA and MCTato Temporal Reasoning tasks, respectively. Note that in these evaluations, demonstrations are picked **randomly**, regardless of their concepts.

significantly smaller volumes and complexity of the training dataset, CoAT-trained models show competitive results to similar-size or even larger in-context learners of previous work. For instance, the 1-billion-parameter Tk-CoAT performs better than the 3-billion T0 in 3 cases (Ax-b, RTE, COPA) and comparably in another 3 cases (WSC, CB, WiC). In comparison with Tk-instruct of the same size, Tk-CoAT-1B outperforms Tk-instruct in 3 out of 7 unseen tasks (WSC, CB, ReCoRD), and reaches similar scores in most other cases, even in 2 out of 3 tasks that were included in Tk-instruct's training mix. Similarly, larger Tk-CoAT-3B outperforms Tk-instruct on 4 of 7 new tasks (Ax-b, WSC, WiC, ReCoRD), but with larger gaps on the others.

C.2 Natural-Instructions: other task types

Figure 8 evaluates the impact of CoAT's mechanism on the quality of in-context learning separately on the English and non-English tasks. The figure reveals that CoAT works particularly well for non-English tasks. Our analyses found this is mainly due to the low performance of the baseline on the non-English tasks. We speculate that this can be a consequence of the higher reliance of the baseline on token semantics (Section 4.6, RQ2); As our models are fine-tuned on an English-only QA model, such learnt reliance is not applicable in multilingual settings.

Figure 9 compares the performance of CoAT models against the models of previous work, separately on the English and non-English tasks. We can see that CoAT is slightly better at the multilingual portion of Natural-Instructions, but the difference is not principal.

C.3 Per-concept evaluations

Figure 7 evaluates the performance gains of the baseline models (§4.2) and CoAT-trained models individually per each of the concepts of the natural datasets. While the CoAT models are able to benefit from concepts the largest in the relative change of quality, they are also not consistent in the ability to benefit from all the concepts. However, as discussed in Section 4.4, this does not imply that CoAT is unable to utilize these concepts.

C.4 Evaluation tasks and other configurations

SuperGLUE (Wang et al., 2019) consists of the following tasks (as ordered in our Results, §4.6): Winogender Schema Diagnostics (AxG) (Rudinger

	# train tasks	AxG	Ax-b	WSC	CB	RTE	WiC	ReCoRD	BoolQ	COPA	MultiRC
Flan-1B	1,836	84.8±3.9	21.9±4.0	70.7±4.8	92.5±2.8*	92.1±3.0*	69.9±5.1*	38.9±5.2*	92.3±2.7*	97.8±1.5*	88.3±3.2*
Flan-3B	1,836	95.3±3.7	22.0±8.0	80.2±9.2	92.7±6.7*	96.0±4.0*	79.7±8.3*	62.2±9.7*	92.1±5.1*	99.3±1.6*	90.4±6.4*
Tk-Instruct-1B	1,616	51.9±4.9	57.2±5.8	49.8±4.9	46.0±5.5	55.5±4.8	53.5±5.3	13.1±3.7	63.4±3.4*	76.9±3.2*	62.2±5.1*
Tk-Instruct-3B	1,616	53.5±4.7	49.9±4.9	51.2±4.9	66.3±4.6	62.7±4.6	50.4±4.8	18.6±4.2	68.8±4.4*	73.8±3.5*	59.9±4.9*
T0-3B	35	65.0±4.5	36.1±4.6	53.5±5.2	48.0±5.4	51.3±5.2	54.0±5.0	20.5±4.0	60.1±4.9	56.8±3.6	56.2±4.4
Tk-CoAT-1B	2	50.4±5.3	52.7±4.6	53.6±5.2	46.9±4.9	53.7±4.9	53.5±5.3	17.0±3.5	63.8±5.4	76.1±3.2	11.4±2.6
Tk-CoAT-3B	2	57.9±4.9	57.2±4.8	53.6±4.5	60.4±4.8	52.0±5.4	56.9±5.0	23.1±3.8	63.6±4.3	81.3±3.3	56.9±3.6

Table 5: **ICL performance: comparison to previous ICL models** ROUGE-L of CoAT-trained ICL models and models of comparable size in previous work. Evaluation setup is consistent with Table 1. In cases marked with *, the task was used in the model's training; Underlined are the best results per unseen task and model size.

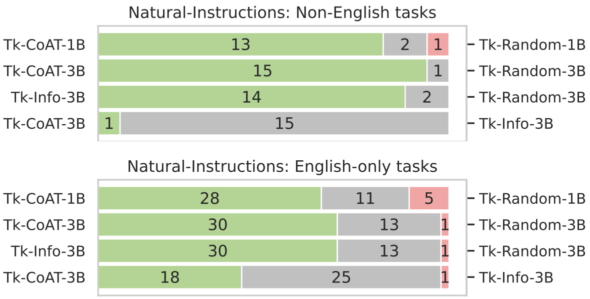


Figure 8: **Impact of Concept-aware training per different language settings:** Pairwise comparison of models trained using selected training configurations (§4.2) on (top) *Non-English* tasks and (bottom) *English-only* tasks of Natural-Instructions collection. Values in green and red bars indicate a number of tasks where the referenced model reaches significantly higher accuracy than the other. For the tasks denoted as *similar*, the difference in performance falls within the evaluation's confidence intervals.

et al., 2018), Broadcoverage Diagnostics (CB), The Winograd Schema Challenge, CommitmentBank (CB), Recognizing Textual Entailment (RTE), ContextWords in Context (WiC) (Pilehvar and Camacho-Collados, 2019), Reading Comprehension with Commonsense Reasoning (ReCoRD) (Zhang et al., 2018), BoolQ (Clark et al., 2019), Choice of Plausible Alternatives (COPA), Multi-Sentence Reading Comprehension (MultiRC).

Natural-Instructions consists of a larger mixture of tasks, which we do not enumerate here to maintain readability; the full list of evaluation tasks can be found in the original work of Wang et al. (2022) in Figures 11 and 12.

To maintain comparability of evaluations among models, we deterministically fix the demonstration selection procedure so that only the full prediction prompts for all the models are the same. In the analyses comparing the differences in performance

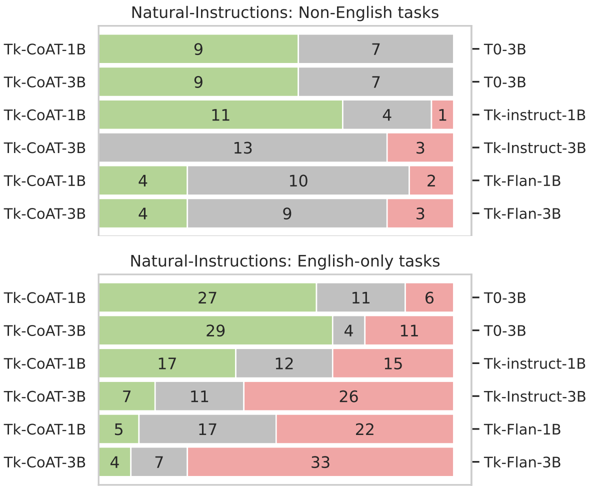


Figure 9: **Comparison to previous work per different language settings:** Pairwise comparison of CoAT models vs. the models of previous work on (top) *Non-English* tasks and (bottom) *English-only* tasks of Natural-Instructions collection. Values denote the number of tasks where the model reaches significantly better accuracy. For the tasks denoted as *similar*, the difference in performance falls within the evaluation's confidence intervals.

(§4.4; RQ1+2), we fixed the prediction samples (x_{pred}) between different demonstrations' sampling strategies to avoid perplexing our comparison with possible data selection biases. Further details can be found in the referenced implementation.

D Computational Requirements

We run both training and evaluation experiments on a machine with dedicated single NVIDIA A100-SXM-80GB, 40 GB of RAM and a single CPU core. Hence, all our reproduction scripts can run on this or a similar configuration. Two stages of training in total take at most 6,600 updates and at most 117 h of training for Tk-CoAT to converge.